Supplementary Results for Roth et al., PyClone: Statistical inference of clonal population structure in cancer

1 Simulated data results

To systematically assess the performance of different modelling strategies for mutational clustering and cellular prevalence inference, we generated 100 synthetic datasets of 100 mutations with randomly assigned copy number and mutational genotypes, grouped into eight clusters in each run. We evaluated the performance of PyClone on these data using five different strategies for specifying the state priors plus two standard clustering models (IGMM, IBMM) outlined in **Online Methods**. Benchmarking was based on V-measure and mean error in estimating cellular prevalence (see **Online Methods**). Group distributions over accuracy were assessed using ANOVA tests with pair-wise TukeyHSD tests for two-group comparisons. Adjusted p values with q<0.05 was used as a criteria for statistical differences. Details of synthetic data generation, the PyClone model and its variants, and statistical analysis is provided in the **Supplemental Note**.

Clustering accuracy was highest in the PyClone PCN method with a V-measure of 0.78±0.06, followed by PyClone TCN (0.65±0.07) (**Supplementary Fig. 12a**). The PCN and TCN methods, which account for mutational genotype, significantly outperformed all other methods (**Supplementary Fig. 12a**). Accounting for copy number but ignoring mutational genotype, the NZ method (0.56±0.06) was worse than the PCN and TCN methods, but performed significantly better than the AB and BB methods that assume diploid states for the variant population (0.49±0.05 and 0.52±0.04, q<0.05, ANOVA and TukeyHSD). The IBMM (0.51±0.05) performed similarly to the PyClone diploid methods but was significantly better than IGMM (0.20±0.11).

Mean error on cellular prevalence estimates was also measured for each method. Similar to V-measure benchmarks, PyClone PCN was the most accurate (mean absolute error=0.03±0.01), significantly lower than all other methods (**Supplementary Fig. 12b**). TCN (0.07±0.02) and NZ (0.14±0.03) were more accurate than the AB and BB methods (0.21±0.03 and 0.20±0.04). All PyClone methods were significantly more accurate than the IBMM and IGMM methods (0.25±0.05 and 0.26±0.05). The likely reason is that the PyClone methods account for tumour content (set at 0.75 for all simulations). Of the two methods which consider mutational genotype, PCN significantly outperformed TCN. Though not surprising since the PCN method provides more informative prior information, this result suggests that given reliable parental copy number information, the PCN strategy for setting genotype priors would improve inference.

Taken together, these results systematically demonstrate the theoretical basis for estimating mutational genotype by incorporating copy number and parental allele information. This confers increased accuracy in both clustering and cellular prevalence estimates. Furthermore these results validate the basis for avoiding the use of Gaussian distributions when clustering deep digital sequencing data.

To asses how performance varies with the number of mutations, we simulated datasets with 50, 100, 500 and 1,000 mutations. 100 replicates were generated for each number of mutations using the same procedure as above. The datasets were analysed using PyClone with Bin-PCN model. We find that clustering performance deteriorates (**Supplementary Fig. 13a**) as the number of mutations analysed increases. In contrast the estimated cellular prevalence of the mutations improves as we increase the number of mutations analysed (**Supplementary Fig. 13b**). We believe that clustering performance deteriorates with more mutations because the problem of clustering larger datasets is intrinsically more difficult. The increasing accuracy of cellular prevalence estimates would suggest that mutations from the same cluster are being placed together. As we increase the number of mutations, the mean number of predicted clusters increases from 8.64 ± 2.02 with 50 mutations to 28.85 ± 13.77 with 1,000 mutations. Taken together this suggest PyClone tends to over cluster as the number of mutations increases, but the cellular prevalence of each cluster will be accurate. This differs from the over clustering of genotype naive methods, which results in mutational clusters being assigned inaccurate cellular prevalence estimates.

2 High grade serous ovarian cancer results

Case 1 mutations clustered by IBBMM (**Supplementary Fig. 14a**,**b**, **Supplementary Table 2**) resulted in mutations with the highest allelic prevalences (Cluster 4, n = 9 mutations) grouped together. These mutations showed a similar prevalence pattern to mutations in Cluster 5 (n = 31 mutations) across samples. We suggest these mutations belong to the same clone, with Cluster 4 mutations predominantly homozygous and Cluster 5 mutations predominantly heterozygous (**Supplementary Table 2**). Eight of nine mutations in IBBMM Cluster 4 fall into regions of heterozygous deletion in all four samples. By contrast, 28 of 31 mutations in Cluster 5 are in diploid heterozygous regions (**Supplementary Table 2**) in all four samples. Therefore, the difference in cellular prevalence estimates in IBBMM between Cluster 4 and 5 can be explained by the copy number of the loci spanning the mutations impacting the mutational genotype. PyClone instead groups the mutations corresponding to IBBMM Cluster 4 and Cluster 5 into one group of n = 44 mutations (**Supplementary Fig. 14c**,**d** - Cluster 1), with representation of both the heterozygous and homozygous loci at cellular prevalences of near 1.0.

PyClone cluster 1 in both case 1 (n = 44 mutations) and case 2 (n = 25 mutations) (**Supplementary Fig. 14d**, **Fig. 2d**) likely represent mutations comprising the ancestral clone in these tumours' aetiology. In contrast, clusters with cellular prevalences lower than 1.0 indicate putative descendant clones. We emphasize that IBBMM splits the ancestral clone into at least two clusters in both case 1 and case 2, with evidence from the copy number analysis (**Supplementary Table 1** and **2**) attributing the split based on heterozygous or homozygous mutational genotype. PyClone modelling of mutational genotypes coupled with simultaneous inference across multiple samples therefore provides a more robust approach to ascertaining clonal populations with dramatic implications for how cellular prevalence estimates of mutations are interpreted in reconstruction of evolutionary histories.

3 Stability of predictions using different copy number predictions

In order to quantify the impact that different copy number predictions have on the PyClone prediction, we analysed two HGSOC cases using three different copy number methods to predict tumour content and inform the PyClone mutational genotype priors (**Supplementary Note**). We ran three random starts of the MCMC analysis for each method to ensure the variability we observed in output was due to different inputs rather than stochastic convergence issues.

Slightly different clusterings and cellular prevalence estimates are observed (**Supplementary Fig. 5**). The difference in clusterings usually result in a single mutation being removed from a large cluster and forming a singleton cluster. The posterior pairwise clustering probabilities largely converge to the same values (**Supplementary Fig. 6 - 11**). These results suggest PyClone is somewhat sensitive to the copy number predictions used to inform the prior. However, this analysis was based on using the PCN strategy for setting mutational genotype priors. The PyClone model allows for flexible priors to be used and we are investigating approaches to combine copy number predictions from multiple methods to improve robustness. In the interim a reasonable practice maybe to remove positions which lie in regions with very different copy number predictions between methods.